UNIVERSITY OF TARTU

Faculty of Social Sciences

School of Economics and Business Administration

Kateryna Volkovska

MODELING THE PREDICTIVE PERFORMANCE OF CREDIT SCORING BY LOGISTIC REGRESSION AND ENSEMBLE LEARNING

Master's thesis

Supervisor: Jaan Masso

Tartu 2018

Contents

1. Introduction 3
2. Literature review 6
3. Methodology 14
   3.1. Random Forest estimation 14
   3.2. Weight of evidence (WOE) and information value (IV) 15
   3.3. Regression analysis 16
   3.4. Scores calculation 17
   3.5. Model evaluation 20
4. Data 21
5. Results 25
   5.1. Random Forest estimation 25
   5.2. Logistic regression output and pre-analysis 27
   5.3. Scorecard evaluation 30
6. Conclusions 33
7. References 35

1. Introduction

Nowadays, an accurate credit scoring is an essential tool for every lending organization and bank institution. Loan market increases rapidly in both size and count of loans (Galindo et. al, 2000) and therefore it becomes impossible for a limited number of individuals to estimate the risk of default of every application manually[1]. Market expansion and the exponential growth of data available has triggered a desire for automatic and fast risk assessment, and consequently, mathematics and statistics serve for building an accurate tool needed for the decision making based on objectivity and scientific principles. Modern machine learning and data mining algorithms contribute significantly (Galindo et. al 2000) to the area of information assessment as they are able to build models that help to estimate the risk level of each customer based on his or her personal characteristics and classify him or her as having a probability of "good" or "bad" (defaulter) according to the perceived risk level. The benefits for such automatic tool are obvious: it saves a lot of time in underwriting which increases the acceptance rates and revenue for the organization as well as reduces salary expenses for credit specialists that perform manual risk assessment. Also, statistical models are significantly more reliable than the manual judgemental assessment by credit specialists as human mistakes are very likely to appear (Steenackers et. al, 1989). The credit specialists have to access manually thousands of applications every day and due to human capacity, it is not possible to fully concentrate at each separate one, take into account all application and external characteristics and make an accurate decision.

It is very important to determine which variables are strong predictors of the customers' ability to repay the granted credit. Factors such as employees' blood sugar (Kuhn et. al, 2014) and the order with which the loans were applied (Chen et. al, 2016) affect which loans are issued, which does not actually have any significance on customers' probability of default. Manual judgement can also lead to the smaller acceptance of less "perfect" customers, who could still be profitable. Thus, there arises a problem of misclassification: the credit specialist might accept the "bad" customers who will default on the granted credit and lead to the loss for credit institution and reject the "good" customers who will appear to be profitable for the organization. The

[1] According to AMR (Allied Market Research), global lending market, valued at $26,064 mln in 2015, is expected to reach $460.3 bln by 2022 (AMR, 2017)

misclassifications (wrongfully accepted "bad" applicants or wrongly rejected "good" ones) can cause massive losses or foregone profits to the lending organization or bank institution and wrong credit policies may even lead to its bankruptcy and collapse. (For instance, 2007- 2008 NINJA Loans (No Income No Job) loans for mortgages contributed to the financial crisis (Hull, 2009).

As the technological advancements have made data analysis easier and much cheaper than it was in the previous century (Mays, 2017), the risk assessment based on statistical models is becoming more and more attractive around the world. Scientists have been analysing different methods and instruments that can be a base for a good risk assessment model, but it is remaining an open question which model is the best tool for predicting consumers' probability to repay.

Usually a credit scoring modelling is performed on the dichotomous dependent variable which assigns "0" to failed loans and "1" to non-failed loans. By default, one means by that variable the inability to repay interest and principal on a loan before the predefined fixed deadline. A credit model is usually constructed on the historical data of the organization and credit bureaus and then is applied on the "live" data to give the decision of granting the loan, i.e. it estimates the odds of repaying the granted credit based on the set of input features (Goovaerts, 1989).

While most of the studies have been concentrated on the development a credit scoring model based on the logistic regression with the dependent variable representing the default rate as described in the previous paragraph (Abdou, 2011), there are fewer approaches that are dedicated to the implementation of machine learning techniques to the credit risk assessment. The few ones that can be found in the literature are Luo et al (2009), Lee et al (2002), Hsieh (2005), etc. Also, there are no approaches that are based on modelling alternative to default-oriented dependent variables.

In the thesis we contribute to the emerging literature by introducing a new target variable based on the surplus, not on the customers' default metric as widely used in the literature (Steenackers, 1989) (Desai et.al, 1997). Such dependent variable is focused not on the default or not default but on the flow of monetary payments from the particular customer. Thus, the customer who is defaulted still can be recognized as "good" by a new metric if the sum of monetary payments from him or her exceeds the amount of issued credit.

Firstly, we define what the surplus is. Surplus is the company's pure revenue, total money collected from the individual minus the amount of granted credit. By total money we mean all money the credit institution collected from the customer: interest, loan sum, fines, extension fees, activation fees, etc. Such surplus-oriented dependent variable was chosen, as interest rates are being squeezed to smaller and smaller values by European central banks[2], and legislation, alternative incoming payments become of significant value (Caballero et. al, 2008) (Pagés et.al, 2016).

Overall, the main purpose of my master thesis is to develop an accurate classification model based on the surplus-oriented dependent variable, rather than the widely used one in the literature that models the standard consumers' default rate. Another novelty and contribution of my master thesis in the context of the existing literature is that for the modelling I do not use only standard logistic regression (Steenackers et. al, 1989) (Finlay, 2010), but an ensemble learning algorithm - Random Forest, that will help me to identify relative importance of the variables and understand which of them are the best predictors of consumer default. Although Random Forest is one of the most common machine learning methods, so far it has given only a couple of contributions to the credit scoring literature. (Sharma et. al, 2010)

While the logistic regression is mostly used for the scorecard development, Random Forest is successfully being applied for prediction in various industries, such as bioinformatics and medicine, finance and banking, stock market exchange, marketing and e-commerce. Random forest has been rarely used for credit scoring, due to business requirements on auditability of understanding around each risk decision, should the need arise. Also, Random Forest is so far unproven technique in banking, and few companies are willing to take a leap of faith, to achieve a small increase in prediction (Sharma et. al, 2010).

The rest of the thesis is structured as follows: in the literature review I summarize the existing papers on the related topic as well as outline some of the most popular existing scorecards. In the methodology part I describe methods used in thesis. Data description part summarizes the data used for the research and in the results, I present the obtained findings and compare them with the previous studies. Section on conclusions summarizes the main outlines of the master thesis and gives suggestions for further research.

[2] Euro area bank interest rate statistics: January 2018. European Central Bank, Press Release 5 March 2018
Euro area statistics: https://www.euro-area-statistics.org/bank-interest-rates-loans?cr=eur&lg=en

2. Literature review

The aim of the scoring card is to provide a company with a fast, stable and reliable way to assess the risk level of the credit application and based on this decide whether to issue the loan or not. The scorecard is based on the statistically significant variables that are used to separate between "good" and "bad" applications. "Bad" accounts are determined as those having the highest probability to default, e.g. not to pay back at least 100% of the borrowed amount.

One of the clearest definitions of credit risk has been given by Zenios (Zenios, 2005, p. 23):

"The risk of an unkept payment promise due to default of an obligor – counterparty, issuer or borrower – or due to adverse price movements of an asset caused by an upgrading or downgrading of the credit quality of an obligor that brings into question their ability to make future payments."

To perform credit-risk assessment, banks and credit institutions usually use scorecard modelling, which by processing the characteristics from historical data of the individuals estimates the chances of the particular applicant to default as well as estimates the expected profitability of lending to the particular group of borrowers with same characteristics (Finlay, 2010). Credit scoring allows lenders to distinguish between "good" and "bad" applications and make statistically sound decision about issuing or not issuing the loan. As the primary aim of the scoring is to provide the lender the best help in minimizing the risk of default, the model can also be used for calculation of the maximum amount that can be lent at the acceptable level of risk for the borrower. In some case it is best to offer to the client smaller quantity of credit, while for the most creditworthy customers lender can decide to offer bigger loan sums.

Credit scorecards became widely used after 1980s (Abdou et.al, 2011). Before, both in banks and the other credit institutions, the most popular way of deciding whether to grant the loan was the humanly-judgemental method, when the risk analyst reviews every application manually and makes his or her decision. It slows the process tremendously as well as increases the probability of biased decision: "bad" applications can be approved, while "good" one rejected.

Nowadays the most popular scoring is the FICO score - it is the gold standard in the consumer credit world (Arya et.al, 2013). First it was developed in 1956 in the United States by Bill Fair and Earl Isaac and later called Fair, Isaac, and Company (hence the acronym is FICO). The score is based on consumer credit information taken from the three biggest world's credit bureaus (external providers): Experian, Equifax, and TransUnion. Clearly, the accuracy of the score is determined by the quality of the information provided. By analysing the information and patterns of past credit reports, FICO determines individual's risk of default.

The scale range of the FICO score is between 300 (minimum possible score) and 850 (maximum possible score). The maximum score is rarely obtained, it demands having a combination of good credit accounts and maintaining an excellent payment history. Usually the range where the most applicants fell is between 600 and 800. The main reasons of receiving low scores are missing payments, debts or bankruptcy.

FICO score bases the decision on analysis of five main categories of information:

- Payment history. The data about how the customer is paying his or her daily bills is proved to be a predictive factor in determining the probability of repaying the credit. Delays and missed payments - lower significantly the chances of obtaining good score and thus receive the loan.
- The amount of outstanding debt. The amount of money one owes, e.g. loans for car purchases, mortgages, as well as the total available credit are strong predictors of the creditworthiness of the applicant. The more money the individual owes the less likely he or she will receive a new credit.
- Credit history data, past credit behaviour of the applicant. The longer is the credit data available per customer's the more chances he or she has in favour of getting new credit. The lender can use the information about past credits and have a clearer picture of the customer attitude to the debt. Individuals paying their debts without delay and not having lots of open loans at the same time are usually the best customers to grant the loan.
- The information about the type of credit. If the individual has different types of credit (mortgages, instalment loans or car loans) he or she is more attractive to the lender than those having e.g. five car loans at the same time as they are probably using the money more wisely and more likely will not default on issued loan. Also, the lender gets more exogenous information about the borrower if he or she has different types of credit.

- The number and frequency of credit applications. The individuals who apply for several times for credit in one period of time are considered to be riskier compared to the others and as the consequence their chances of getting new credit must be reduced.

The most reliable and accurate way to make credit decision is not only to trust the existing scores (Fico, Equifax, Experian) but to analyse the output scores from different sources as well as to make an own assessment (the company's internal scoring card, (Shannon, 1948). A serious drawback of relying only on the existing scorecards is that if the individual does not use credit, he will not get a credit score at all, as there is no stored data about this applicant. Nonetheless people avoiding credit often have the lowest credit risk, as they are saving money for example for other big purchases. On the author's opinion, a lender should always use the own underwriting model, as to the alternative to scorecards based only on external data, that also considers customers' university/education, employment, income, marital status, number of children, etc.

Logistic regression is considered to be the most easily interpretable and widely used in scorecard development (Desai et.al, 1997). Despite of that there have been also lots of attempts in science to use the machine learning algorithms and combinations of different models to build a more accurate scoring algorithm. Many classification algorithms used for scorecard development can be found in the existing literature. Most commonly used ones are the following (Donga et.al, 2010):

- statistical models (logistic regression techniques, linear discriminant analysis, k-nearest neighbour, classification tree);
- operational research methods (linear programming, quadratic programming);
- artificial intelligence techniques (neural networks, support vector machines, genetic algorithm and genetic programming);
- hybrid approaches (fuzzy systems and neural networks, fuzzy systems and support vector machines, neural networks and multivariate adaptive regression splines);
- ensemble models (neural network ensemble, Random Forest).

Luo et. al (Luo et.al, 2009) proposed using support vector machines (SVMs) - clearly machine learning classification algorithm - and tried clustering-launched classification models in their paper. Lee (Lee et al, 2002) have proposed using neural networks with linear discriminant

analysis in credit scoring. Hsieh (Hsieh, 2005) used a hybrid data mining algorithm in the construction of the scoring model, he also employed clustering algorithms and artificial neural networks.

While the above discussion in all cases considered the forecasts of one particular model, one possibility to improve the forecasting accuracy is to combine the forecasts of different models. The main purpose of the combinations of forecasts' is to use a unique variable, that is uncorrelated to the other variables in the dataset, in each classification model to capture as specifically as possible different non-linearities[3] in the data which boost the classification accuracy. First, Bates and Granger (Bates et.al, 1969) proved that a linear combination of different models would give a higher predictive power.

One of the first studies that used various machine learning algorithms for credit scoring was Yao (2009). He used three pre-modelling strategies in his paper:

- classification and regression trees to determine the most predictive input variables for the further modelling;
- multivariate adaptive regression splines for the enhancement of the clear specification of the risk-drivers;
- genetic algorithms for parameters' optimization.

Among other scientists that have been trying to apply machine learning algorithms for the credit risk assessment were Bijak and Thomas. They optimized in parallel a segmentation and scorecard development by using Logistic Trees (Bijak, 2012).

Khashei and Hamadani (Khashei, 2012) were trying to boost the accuracy of the hybrid classification model by using traditional multi-layer perceptions - the simplest neural network[4] prototype. West (West, 2000) compared the performance of several neural network models with traditional techniques such as linear regression and discriminant analysis. His results proved that neural network can improve significantly the accuracy of credit scoring and become a sufficient alternative technique in construction of the credit scoring model.

Baesens et al (2003) performed a wide comparison involving several machine learning algorithms: discriminant analysis, linear programming, support vector machines (SVM), neural networks, Bayesian networks, decision trees and k-nearest neighbour (k-NN). The authors

[3] Nonlinearity is a relation between data points that can't be condensed into a neat linear graph (Cottle, 2017).
4 Henley and Hand (Hand, 1997) define neural networks as: "A statistical model involving linear combinations of nested sequences of non-linear transformations of linear combinations of variables" (pp. 534).

concluded that machine learning methods, if applied accurately, give much higher accuracy compared to basic logistic regression.

Next table is aimed to provide a short and structured summary (which methods the authors used, data description and key results) of the studies that contributed to the literature on introduction of machine learning algorithms to credit risk modelling.

Table 1. Summary of literature review.

Author and paper	Sample and variables	Methods	Results
Luo et al (2009)	The German and Australian credit data sets from the UCI Repository of Machine Learning databases. The German credit data set consist of a set of loans containing a total of 1000 applicants. For each applicant, 20 input variables describe the credit history, account balances, loan purpose, loan amount, employment status, and personal information. The Australian credit data set consist of a set of loans given to a total of 690 applicants.	Support vector machines (SVM) and clustering-launched classification (CLC) models.	CLC overperform SVM, therefore, CLC is an effective tool to construct credit scoring model and should be used by banks and microfinance institutions.
Lee et al (2002)	The dataset by a local bank in Taiwan which consists of 9 predictor variables: personal data variables, educational and occupation related variables, annual income, residential status and credit limits. 6000 datasets with respect to the ratio of good and bad credits were randomly selected and then used to build credit scoring models.	Neural networks with linear discriminant analysis.	Neural network model has the highest average correct classification rate in comparison to discriminant analysis and logistic regression as well as has better capability of capturing nonlinear relationship among variables. Designed hybrid model has not only better credit scoring accuracies, but also has the lowest Type II error associated with high misclassification costs.
Hsieh (2005)	German and Australian credit data sets from the UCI Repository of Machine Learning Databases. The German credit data set consisted of a set of loans given to a total of 1000 applicants, 700 samples of creditworthy applicants and 300 samples where credit should not be extended. For each applicant, 20 variables described credit history, account balances, loan specific characteristics, employment status, and personal information. The Australian credit data set is a similar data set with 690 samples, in which 468 samples were accepted and maintain good credit and 222 samples were accepted but became delinquent.	Hybrid data mining algorithm, clustering algorithms and artificial neural networks.	Hybrid approaches perform very well and have shown better forecasting performance than those of any individual methods.
Bates and Granger (1969)	The 1001 time series used in Makridakis et al. (1982, 1983).	Combined forecasts by using weighted averages.	A linear combination of different models would give a higher predictive power then a single model.
Yao (2009)	Australian and German datasets available from the UCI Repository of Machine Learning databases and are adopted to evaluate the predictive accuracy. Same as in Hsieh (2005).	Classification and regression trees to determine the most predictive input variables for the further modelling;	Hybrid model has the best overall classification accuracy and also the lowest type-II error.

		multivariate adaptive regression splines (MARS) for the enhancement of the clear specification of the risk-drivers; genetic algorithms for parameters optimization.	
Bijak and Thomas (2012)	Data was provided by two of the major UK banks and one of the European credit bureaus and contains application data and behavioural (credit bureau) data.	Optimized in parallel a segmentation (dividing the population into several groups and building separate scorecards for them) and scorecard development by using Logistic Trees (CART)	Segmentation does not always improve model performance in credit scoring. It is recommended to develop a single-scorecard model for comparison purposes.
Khashei and Hamadani (2012)	The data sets include a synthetic data set by Ripley (1994) and real-world data set of diabetes diagnosis among Pima Indians (Asuncion & Newman, 2007). The Ripley synthetic data set is created by Ripley (Ripley, 1994). The data set consists of 1250 samples with two attributes. The two classes are equally represented in the data set.	Boosted the accuracy of the hybrid classification model by using traditional multi-layer perceptron - the simplest neural network prototype.	Hybrid model exhibits effectively improved classification accuracy in comparison with traditional artificial neural networks and also some other classification models such as linear discriminant analysis, quadratic discriminant analysis, K-nearest neighbour, and support vector machines.
West (2000)	Australian and German credit datasets. Variables account longevity, credit history, employment classification, checking account status, assets owned, years in residence, other existing loans, housing classification, loan-specific characteristics, years employed, and savings account status.	Compared the performance of several neural network models with traditional techniques such as linear regression and discriminant analysis	Neural network can improve significantly the credit scoring accuracy and become a sufficient alternative technique in construction of the credit scoring model.
Baesens (2003)	The data sets Australian credit and German credit from the UCI Library (Lichman, 2013) and the data set from Thomas, Edelman, and Crook (2002). Three other data sets, Bene-1, Bene-2, and UK, were collected from major financial institutions in the Benelux and UK, respectively. Data is pooled based on the same product and time period. Datasets capture information from the application form (e.g., loan amount, interest rate, etc.) and customer information (e.g., demographic, social-graphic, and solvency data)	Performed a wide comparison involving several machine learning algorithms: discriminant analysis, linear programming, support vector machines, neural networks, Bayesian networks, decision trees and k-nearest neighbour.	Accurately applied machine learning algorithms give higher accuracy in comparison to logistic regression.

The results of the existing scarce literature on the machine learning applications for credit scoring motivate further looks into this topic. Additionally, as those studies are based on modelling the default-oriented variable, the author also decided to develop the model on another response feature and assess its predictive performance.

3.Methodology

3.1.Random Forest estimation

To check the relative performance of the independent variables I employ the machine learning algorithm called Random Forest. It is a powerful tool capable of delivering performance that makes it to be among the most accurate methods to date. Random Forest algorithm ranks variables based on their predictive ability.

Random Forest (introduced by Breiman in 2001) is an ensemble learning method for classification and regression that constructs a set of separate models (called individual decision trees) and as the result classifies the object according to the mean (average) predicted outcome of those models (Breiman, 2011). Random Forest is proved to overperform many other machine learning classifiers, such as discriminant analysis, neural networks, support vector machines. It is also robust against overfitting (Svetnik, 2003). Given Random Forest is constructed on many trees (individual models) the error rate is also expected to be small (Bylander, 2002) and, at the same time, estimates of variables' importance are becoming more stable (Breiman, 2011). So, the performance of the algorithm is better with the larger number of predictors.

Random Forest is more accurate than the individual classifier as it compiles inside a lot of single independent weak models. Each tree classifies the object, based on its characteristics, in the certain class, in the terminology of public economics, the tree "votes" for the class. The Forest makes the final decision by choosing the class that got the most votes over all trees.

Random Forest is also useful in helping to understand the importance of variables[5] that can be potentially included in the model. The importance of variables measured as the decrease in Gini that shows the average benefit of purity by splits of each feature. For the given variables, it tends to split mixed labelled nodes into single nodes. Those splitting's by a permuted variable should neither to increase nor decrease node purities, that are representing how well the trees are able to split the data. Permuting a useful variable is giving larger decrease in mean Gini gain. Thus, the variable has higher mean decrease in Gini and situated on the top of the graph indicating the more useful variable (Figure 1). Understanding the importance of variables is very

[5] The algorithm estimates the importance of a variable by looking at how much prediction error increases when out-of-sample data for that variable is permuted while all others are left unchanged (Liaw, 2002).

important as it shows which of them have higher predictive performance in relation to the binary outcome and therefore are giving higher weight in the final scoring card.

3.2. Weight of evidence (WOE) and information value (IV)

Accessing the importance of variables is also possible by calculating the information value and weight of evidence measures. Information value is a useful concept for variable selection during model building and scoring card development. The history of information value concept starts in information theory proposed by Claude Shannon (Shannon, 1948). In general, information value estimates how well a variable X can be distinguished between a binary response Y ("good" or "bad"). The idea that is if a regressor has a small information value, it may not be useful for classifying the target variable, and hence it should be removed as an explanatory variable for further modelling.

Weight of evidence (WOE) aimed provide a tool to recode the values in continuous and categorical predictor variables into discrete categories, and to assign to each category a unique WOE value that shows the predictive power of an independent variable in relation to the dependent variable. WOE can also be interpreted as the measure of the "strength" of groupings for separating "good" and "bad" risk (in our context "bad" risk is interpreted as default).

This measure is calculated by taking the natural logarithm of the ratio of the per cent of events occurring and events not occurring. The formula is presented below (Siddiqi, 2005):

$$WOE = [ln\left(\frac{Frequency\ of\ Good}{Frequency\ of\ Bad}\right)] \times 100 \quad (1)$$

The lower the WOE, the higher is the percentage of the "bad" loans, and the opposite, the higher the WOE is the lower is the percentage of "bad" loans. Negative number implies that the attribute is being able to isolate a higher proportion of "bad" than "good" cases. It is important to note that the higher the difference between groups (in terms of WOE), the higher is the predictive ability of the characteristic (variable). All members within one group have the same odds of performing in the certain way.

Another important characteristic for accessing the predictive power of the variable is the information value (IV) that ranks variables based on their predictive ability (Hand and Henley, 1997). According to IV the risk analyst decides which variables into include to the model. The formula for the calculation of that indicator is the following:

$$IV = \sum(Frequency\ of\ Good_i - Frequency\ of\ Bad_i) \times \ln(\frac{Frequency\ of\ Good_i}{Frequency\ of\ Bad_i}) \quad (2)$$

Considering the formula (1), we can rewrite it as follows:

$$IV = \sum(Frequency\ of\ Good_i - Frequency\ of\ Bad_i) \times WOE_i \quad (3)$$

The rules for accessing the power of the variable based on the information value are next (Siddiqi, 2005):

- If IV is less than 0.02, then the predictor is not useful for modelling (in my case - separating the "good" from the "bad") - useless predictor;
- If IV is from 0.02 to 0.1, then the predictor has only a weak relationship to the "good"/"bad" odds ratio - weak predictor;
- If the IV is from 0.1 to 0.3, then the predictor has a medium strength relationship to the "good"/"bad" odds ratio - medium predictor;
- If the IV is 0.3 or higher, then the predictor has a strong relationship to the "good"/"bad" odds ratio - strong predictor;
- If the IV is higher than 0.5 - this means "too good to be true" - suspicious predictor. This is quite rare case and often means the analyst made a mistake in calculations.

3.3. Regression analysis

Logistic regression is aimed to find the model, that best describes the connection between the response and the set of independent variables. Logistic regression is based on odds - the probability that an event will occur (let's define it as *p*) divided by the probability that it will not occur *(1 - p)* The odds is a value given by following formula (Park, 2013):

$$odds = \frac{p}{1-p} \quad (4)$$

Here in the numerator we can see the probability of event occurring (the loan will be defaulted) over the probability of an event not occurring (the loan is defaulted). According to the definition of logistic regression, the "core" of modelling is the natural log of odds as a function of the regressors:

$$logit(y) = \ln(odds) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_i x_i \quad (5)$$

where α – intercept of linear regression, β_i – slope of the linear regression, x_i – explanatory variable and y is the dependent variable.

The basic assumptions of the logistic regression are (Park, 2013): the dependent variable is binary (default/not default in our case), observations are independent between each other (only one person from the household can apply for the credit) and the lack of multicollinearity. Multicollinearity assumes the linear relationship (correlation) between the independent variables in the model. As a result, the estimation becomes biased and insignificant, which leads to incorrect conclusions. For checking multicollinearity, the variance inflation factor is used. It provides an index that measures how much the variance of an estimated regression coefficient is increased because of collinearity (Allison, 1999).

The formula for VIF calculation is presented below:

$$VIF_i = \frac{1}{1 - R_i^2} \quad (6)$$

where R_i^2 is the coefficient of determination of the estimated regression equation for the regressor X_i. If VIF (Variance Inflation Factor) larger than 4.0 for one variable, then there is multicollinearity and this variable should be removed before fitting them into the model (Sheather, 2009).

To build logistic regression model and Random Forest, I split the data into two sets: the training set (80% of the data set) and the test set (20% of the data set). On the training set the model "learns" from the data. This set has the already labelled dependent variable and the model is finding patterns in the data and identifies which characteristics lead to certain outcomes (default and non-default). The test set has only independent variables and the model classifies each observation given the independent variables and the experience it got from "learning" the training set. It is important to verify the out-of-sample forecasting performance of the constructed Random Forest to see how well the model is able to predict the dependent variable.

3.4. Scores calculation

Regression coefficients are used to calculate the scores for each attribute of the variable. For this we will use the following formulas (Siddiqi, 2005):

$$Score = \left(\beta \times WoE + \frac{\alpha}{m}\right) \times factor + \frac{offset}{m} \quad (9)$$

where:

$$factor = \frac{pdo}{\ln(2)} \quad (10)$$

$$Offset = Score - (factor \times \ln(Odds)) \quad (11)$$

and:

- β - coefficient from logistic regression
- α - intercept from logistic regression
- *WoE* - Weight of Evidence for each attribute of the variable
- *m* - number of variables included in the model
- *factor* - scaling parameter
- *pdo* - points to double the odds (usually predetermined by the analyst)
- *score* - shift, the scores will be around it (usually predetermined by the analyst)
- *odds* - odds of the loan repayment for specific scoring value (usually predetermined by the analyst)

Score, factor and odds are used to transform the expected default rate into a user- friendly form. Using the above formulas, we can compute the scores for each characteristic and for each individual variable. Then the final scoring card will look as shown in Table 2:

Table 2. Example of the sub-scoring card for variable 1.

Variable name	WOE	Score
V1		
[a; b)	0.245	27.06
[b; c)	-0.133	20.18
[c, d]	-0.203	18.83

Having an application with the variable $V1 = k \in [a; b)$, will lead to assignation of 27.06 points for this characteristic $k \in [a; b)$, when the variable V1 falls in the range from *a* to *b* and *k* is any number from the interval $[a; b)$. For example, if the variable is age and the applicant is 20 years old, while the scorecard assigns 27 points to the age from 18 to 30, the application will get 27 points for its age characteristics and same procedure is repeated to all other variables.

Every variable included in the model should have own "sub-scorecard" – the allocated scores for its each split. Then all scores are summed up to yield the total score for the application.

Mathematically, the score can be explained as the probability of repaying the loan and it is derived from the set of applicant characteristics. The examples of these characteristics (variables) are gender, age, employment status, marital status, postal code, etc. The older customers are more likely to repay the loan and that's why they are getting higher score. The customers that don't have currently an open loan but have already had several repaid loans are most likely to repay the loan as consequently are receiving higher points. Also, as an example could serve those customers who have relatively small income, but a couple of open loans – they are less likely to receive the next loan as they are likely to face tight financial constraints in the near future.

The company's approval rate is another related definition. It is calculated as the fraction of the loans approved to the total number of loan applications. The company sets the approval rate by many factors, such as minimizing the default rate, increasing portfolio growth, and also by time of year, marketing campaigns etc. In the extreme to approve all customers would be to allow all good and bad into the portfolio. And in the other extreme to remove all bad clients the institution would be required to set approval rate to minimum. The threshold for this in-between the extremes is also called a cut-off. For example, if the company decides to accept all applications that have the score higher 280, then 280 is the cut-off. Formally,

- Application is accepted ⇔ total score for application >= cut-off
- Application is rejected⇔ total score for application < cut-off

To set statistically derived threshold that is optimizing the trade-off between the default rate and sales is usually the most important task for the risk analyst.

3.5.Model evaluation

To evaluate the model, I use Gini coefficient. It is defined as a ratio of the areas on the Lorenz curve[6] diagram (Chen, 2004). If the area between the line of perfect equality and Lorenz curve is *A*, and the area under the Lorenz curve is *B*, then the Gini coefficient is $\frac{A}{A+B}$.

If the Lorenz curve is represented by the function *Y = L(X)*, the value of *B* can be found with integration (Gastwirth, 1972):

$$GINI = 1 - 2 \int_0^1 L(X)dx \quad (12)$$

Gini coefficient always falls in the interval *[0, 1]*, where 1 indicates the strongest scorecard - the scorecard drops all "bad" outcomes and keeps all "good" outcomes. Too high Gini might indicate overfitting. If the Gini coefficient is approaching to zero then the scorecard is weak, it means the model is not able to distinguish accurately between "good" and "bad". In practice with the real data, however, models don't have Gini higher then 0.5, and when Gini is equal 0.3 - it is already considered as a fairly good model.

[6] A graphical representation of the inequality between two distributions (usually used to represent the differences of income or of wealth) (Gastwirth, 1972)

4. Data

Building a scorecard implies a careful and statistically substantiated selection of variables. The independent variables can be divided into two classes: numerical and categorical. The example of a numerical variable can be the number of already repaid loans, the customer can have repaid by now 0 loans as well as 1, 2, 3, …, 100 and more. The example of a categorical variable can be home ownership type. Let us suppose that the home ownership type can take one of three values, these are rent, owner and owner with mortgage. Subsequently, the points in the scoring card would be divided between those three categories that will describe, how likely an applicant, having certain home ownership type, is likely to repay the credit.

Every continuous variable can be represented also as a categorical variable. For example, income is usually used as a numerical variable, while when filling the application customer puts a number representing his income in the corresponding box. However, to reduce mistakes and frauds the lender can make income a categorical variable allowing the applicant to choose only from the list of predetermined answers – splits or ranges. In this case the variable net income becomes categorical variable.

The dataset used in the theses originates from the portfolio of a particular lending organization. It consists of 8268 observations, 13 independent features (variables) and a dependent variable. Each observation is a single loan taken by the customer with particular characteristics. I take the data from the employer's databases and due to information security reasons not able to reveal the names of variables. Hence, the names are represented in the general way.

The descriptions of variables are presented below in table 3:

Table 3. The description of the variables used to build the scoring card.

Variable label	Type of the variable	Source of information
V1	Numerical	From internal database
V2	Numerical	From external provider

V3	Numerical	From external provider
V4	Numerical	From external provider
V5	Categorical	From external provider
V6	Categorical	From application form
V7	Categorical	From application form
V8	Numerical	From application form
V9	Numerical	From application form
V10	Numerical	From external provider
V11	Numerical	From application form
V12	Binary	From internal database
V13	Numerical	From external provider

Source: own elaboration based on the dataset

To build my model I constructed the dependent variable that defines whether the loan is "good" or "bad" in the following way:

- loan is "good" - (the dependent variable takes the value of 0) if the customer paid the financial institution before the final deadline plus 30 days more than 100% of the loan principal amount (surplus is positive or zero). Final deadline is the day, agreed between the applicant and the financial institution during the application process, until which, the customer is expected to pay the granted credit, interest amount and related costs fully;
- loan is "bad" - (the dependent variable takes the value of 1) if the customer was not able to pay the financial institution before the final deadline plus 30 days more than 100% of the loan principal amount (surplus is negative).

We are interested whether the money that the institution lends are coming back in a full amount, however, "bad" loan can still be "open", meaning some payments that will increase

surplus (monetary gain from the customer) will be still expected at least in some cases. So, the future model will be based on the surplus-oriented measure, rather than the standard default rate, meaning we are not predicting whether the loan will be "defaulted" or "not defaulted", but we are focusing whether the loan could potentially bring the monetary gain, rather than the loss for the financial institution. The customer who is defaulted still can be recognized as "good" by a new metric if the sum of monetary payments from him or her exceeds the amount of granted credit.

Variables for scorecard creation are taken from different sources for example credit application form, external credit bureaus, government database, internal and external fraud detection databases. Application data is provided by the customer while filling the application form on the website of the lending company. The bank or microfinance institution should keep the application form simple and easy to understand as it will increase the chance to complete the application and move to the lending system for the risk assessment. The problem with the data from application form is that the applicant can intentionally lie about his characteristics, for example about the income: customer might think that putting high income will increase the chances of receiving the credit or let us take an example with the home ownership type: an applicant might declare owning an apartment while he actually rents the one from the landlord. These situations are typically called as frauds and should be investigated by the fraud analyst in the company.

For the information from credit bureaus the lender should pay, and this information is usually very useful for scorecard development. The examples of the variables are the number of loans customer has paid in the past, the number of days in overdue or number of rejected applications. The drawback using this data in credit scoring is obvious: customers that do not have any credit history would not have the data and would have smaller chances to get the loan, but a lot of those customers could be "good" borrowers. However, there is another complication that those customers which were for a long time not needing credit, their financial situation may have changed so that now they need a credit.

Another source of data that the company might find useful to collect is the information about the customers' past payment behaviour. The past habits of the individual are strong predictors whether the customer is likely to repay the next granted loan. If the customer went 30

days overdue on the one of his loans it is highly likely he will also fail to repay the next loan at due date.

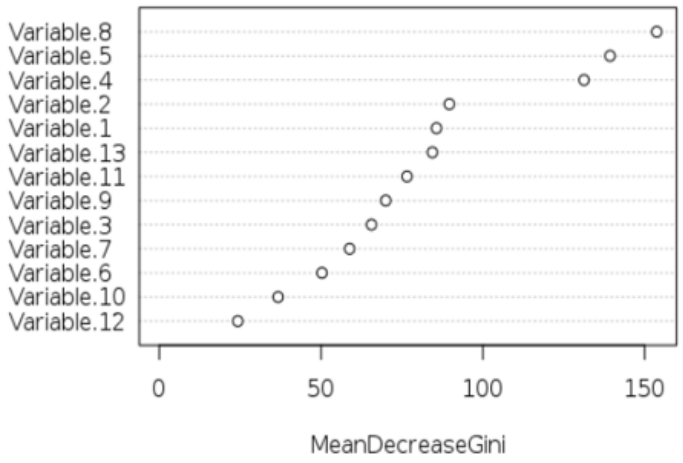
It is important that the scorecard should have a mix of variables from different sources, to avoid it to be too much dependent from one type of information (Mays, 2017) – that is why I am mixing variables from different sources in my model. For example, if the 90% of the variables are from application form then the scoring card can be biased as in the application form the individual can intentionally hide or violate true data. The person can easily misreport his or her income to increase the chance of getting the loan, that is why it is very important for the financial institution to verify the customer income or take the income data from the external agencies. In addition, using external data, makes it harder for the potentially "bad" customers to commit fraud attacks[7].

[7] Planned actions aimed to intentionally harm the organisation's security, e.g. steal or delete the data.

5. Results

5.1. Random Forest estimation

As an output of Random Forest algorithm, I will present standard variables' importance estimation (Grömping, 2012) and the confusion matrix which indicates the quality of constructed prediction model. We start the overview of the results by reporting the importance of variables that we will use for the scorecard construction: the higher the variable is positioned at Figure 1 - the more important it is for the current model, and as a result it is getting the higher weight in the final scoring card and more influence in the further risk assessment.



Source: Private database, author's own calculations

Note: The most important variables are at the top and an estimate of their importance is given by the position of the dot on the x-axis

Figure 1. Variables' importance plot.

On the x-axis is displayed the mean decrease in Gini for each of the variables in relation to the binary outcome – "good"/ "bad" (default/not default). We can notice that the Variable 8 and Variable 5 are appearing to have the highest influence among others, while Variable 10 and Variable 12 are the least influential in this set. Intuitively, this is logical, algorithm found the strongest variables from the two different data sources, from the credit application and the external provider.

My model is relatively accurate; the accuracy is 68% on the test set, meaning the model can relatively well classify the unseen data. I estimated the confusion matrix for my model (Table 4), that is gives an overview how well the model can predict the data. From the confusion

matrix below we can see that 1250 (true negative[8]) + 154 (true positive[9]) = 1404 cases are correctly classified, while 179 (false positive[10]) + 485 (false negative[11]) = 665 cases are classified incorrectly. The model has the accuracy of 67% with a 95% confidence interval - (0.66, 0.7).

Table 4. The confusion matrix: results

	Reference	
Prediction	1	0
1	154	179
0	485	1250

I also built the Receiver Operating Characteristic (ROC) curve for Random Forest that visualizes the predictive strength of the model. ROC curve is the true positive rate (in machine learning also called sensitivity) plotted as function of the false positive rate (in machine learning called specificity) (Akbani, 2004) for different acceptance thresholds. Every point on the ROC curve is a sensitivity-specificity pair that corresponds to each cut-off. The closer the curve is to the left upper-angle - the better is the performance of the model. If the curve coincides with the 45-degrees line, then the model is not better than the random guess.

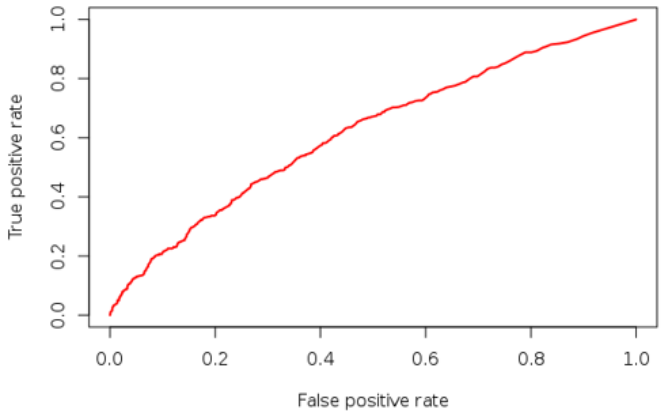


Figure 2. ROC curve for Random Forest.

8 **True negative (TN):** condition is not detected when it is absent.

9 **True positive (TP):** condition is detected when the it is present.

10 **False positive (FP):** condition is detected when it is absent.

11 **False negative (FN):** condition is not detected when it is present.

From the graph we can see that the model has average performance (ROC curve is above, but not too far from the 45-degree line) on the test set. Area under curve of the model is 0.68 - usually considered as fair (Bylander, 2002).

5.2. Logistic regression output and pre-analysis

The linear correlation means the existence of the linear relationship between the variables. Below I am presenting the correlation matrix for the variables from my dataset.

Table 5. Correlation matrix

	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11	Var12	Var13
Var1	1.00												
Var2	0.00	1.00											
Var3	0.00	0.04	1.00										
Var4	-0.04	0.07	0.40	1.00									
Var5	0.03	0.17	0.40	0.39	1.00								
Var6	-0.04	-0.02	0.01	0.04	0.04	1.00							
Var7	0.16	0.00	0.02	0.03	0.06	-0.05	1.00						
Var8	0.03	0.02	0.03	0.01	0.01	0.02	0.04	1.00					
Var9	0.01	-0.02	-0.01	-0.03	-0.06	0.00	-0.12	0.01	1.00				
Var10	-0.03	-0.01	0.34	0.33	0.25	0.04	0.00	-0.02	0.00	1.00			
Var11	-0.18	0.02	0.01	0.05	0.06	0.00	0.00	0.02	0.02	0.00	1.00		
Var12	-0.02	-0.08	-0.04	-0.09	-0.08	0.02	0.00	0.00	0.02	0.08	-0,10	1.00	
Var13	-0.03	0.26	0.17	0.27	0.30	0.01	0.05	0.02	-0.06	0.12	0.03	-0.08	1.00

On this stage of the scorecard development, the risk analyst should investigate the strength of the correlation between the variables and exclude too strongly correlated variables as if to keep them the estimation would result in biased coefficients and the effect of the features in the final scoring card would be double-counted. Usually, by the rule of thumb, the correlation of 0.5 or larger is considered as strong correlation between variables (Tabachnick, 1996). From the presented table we can conclude that there will not be too strongly correlated variables in the model, meaning we do not have to exclude any variables for further analysis and model construction.

Next step is to calculate WoE and information value of the variables. From the table below, we can deduce that my future model will consist from medium and weak predictors that is a typical case on practice (Mays, 2017). We can also draw a strong parallel with the Random Forest variables' importance estimation (Figure 1): variable 5 and variable 4 are the strongest relative to the others and separate "good" account from the "bad" accounts with the highest accuracy.

Table 6. Information values and estimations' output

Variable	# of splits (categories)	Information value	VIF	Coefficients' estimates
Intercept				0.633***
Variable 1	4	3.5%	1.077	1.235***
Variable 2	4	2.1%	1.110	0.649***
Variable 3	4	12%	1.394	0.421***
Variable 4	5	15%	1.427	0.626***
Variable 5	7	11.2%	1.429	0.534***
Variable 6	4	8%	1.014	0.951**
Variable 7	2	10%	1.044	0.101**
Variable 8	6	7%	1.006	0.983**
Variable 9	3	7%	1.223	1.378***
Variable 10	3	6%	1.251	0.290***
Variable 11	4	2%	1.047	0.843*
Variable 12	2	6%	1.048	1.761***
Variable 13	5	5%	1.214	0.451***

Note: p-value < 0.1; ** p-value < 0.05; *** p-value < 0.01

The forth column shows the VIF that is described in the methodology section. As VIF for each variable is less than 4.0, no multicollinearity is detected, and as other assumptions of the

logistic regression are satisfied, one can estimate the model. From the Table 7. we can also see that all coefficients have the same sign and are significant at least at 10% level. We can interpret the coefficients as the power each variable contributes to the final scorecard. The larger the magnitude of the coefficient the more points are between different values of the same variable, meaning the more precisely the variable can separate the potentially "good" applications from the potentially "bad" applications. The coefficients of the logistic regression are later used for the calculation of credit scores.

The equation for my model has the following form:

$$\log\left(\frac{p}{1-p}\right) = 0.633 + 1.235\times V1 + 0.649\times V2 + 0.421\times V3 + 0.626\times V4 + 0.534\times V5 + 0.951\times V6 + 0.101\times V7 + 0.984\times V8 + 1.378\times V9 + 0.290\times V10 + 0.843\times V11 + 1.761\times V12 + 0.451\times V13 \quad (8)$$

I am also presenting the heat map of the expected default per score - the expected average frequency of "bad" outcomes per score interval or, in other words, the expected probability of default. The colours are the conditional formatted numbers - the higher the number - the redder is the particular cell, the lower the number (compared to other numbers in the table) - the greener the cell. Depending from the default rate the financial institution expects to have it chooses the corresponding threshold for accepting the loan application.

Score range	Average expected default
<280	
280-281	
282-283	
284-285	
286-287	
288-289	
290-291	
292-293	
294-295	
296-297	
298-299	
300-301	
302-303	
304-305	
>306	

Figure 3. The expected average frequency of "bad" outcomes per score range

Note: Colours are indicating the relative magnitude: the redder the cell the higher the number (meaning higher expected default), the greener the cell the smaller the number.

In addition, Lorenz diagram can be used to represent how well the model distinguishes between the "good" and the "bad" loans. On the horizontal axis is the proportion of "bad" loans against the proportion of "good" loans on the vertical axis. The diagonal line from the bottom left corner of the chart to the top right corner represents a model that cannot distinguish between "good" and "bad" at all, so the model is not better than the random guess. The better the scorecard is, the larger is the difference between proportions of "good" and "bad" - meaning for the good scorecard the difference between the two lines is significant. Figure 3 below present the Lorenz diagram for the model estimated in this paper.

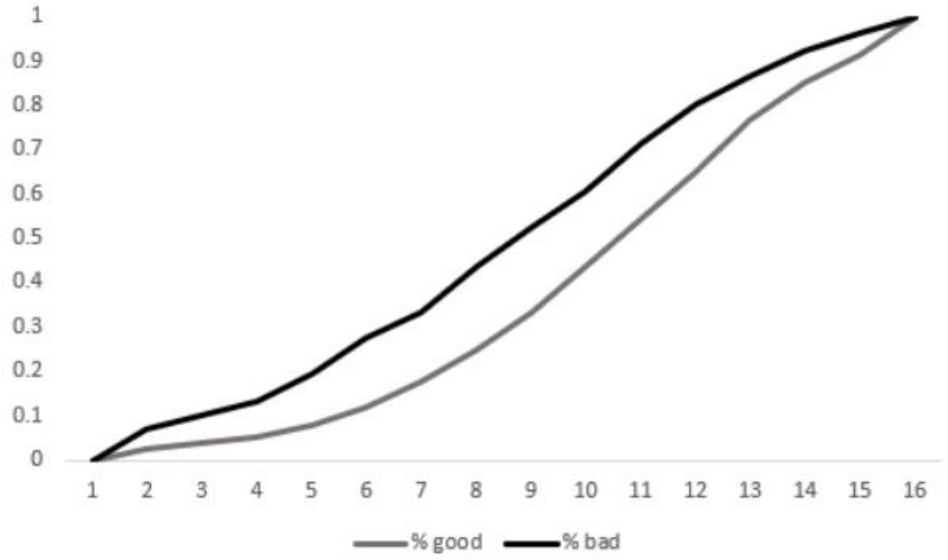


Figure 4. Lorenz curve

Note: On the x-axis the number of score bins (splits) is represented, it is the same as in the above heat map (Figure 3).

From the above graph, we can draw a conclusion that the developed scorecard can distinguish quite clearly "good" accounts from "bad" accounts as there is a significant difference between two curves.

5.3. Scorecard evaluation

The last and one of the most important stages in the scorecard development is to check how well the model performs over time (Siddiqi, 2005). As if, for example, scoring card was constructed on the December data and risk analyst noticed that the December data differs significantly from November and October data, it means that in January the company might not

get the expected results. The scoring card needs constant monitoring to notice if the distribution of incoming independent variables changes over time that should help to avoid undesired biases (Mays, 2017). Sometimes it can happen that the population changes and then it is best to rebuild the scorecard from scratch and re-estimate all the scores. Also, an important part is to consider the changes in economic environment (Zhang and Lyn, 2015), e.g. increase in unemployment or inflation can lead to unexpected results in applicants' performance. Therefore, if the analyst notices any unexpected changes the scorecard should be adjusted immediately and properly.

Before the implementation of the model, the analyst should check that the expected average default rate is stable over the available historic data and if some month differs significantly, then exclude them from the sample and rebuild the scoring card (Siddiqi, 2005). Below is the heat map of the scorecard performance relative to the average probability of default on the historical data (couple of months before the start of scorecard development). The colour formatting is explained in the previous heat map (Figure 3), however, here the main point is to see that colours are in line in all month for the particulate score range.

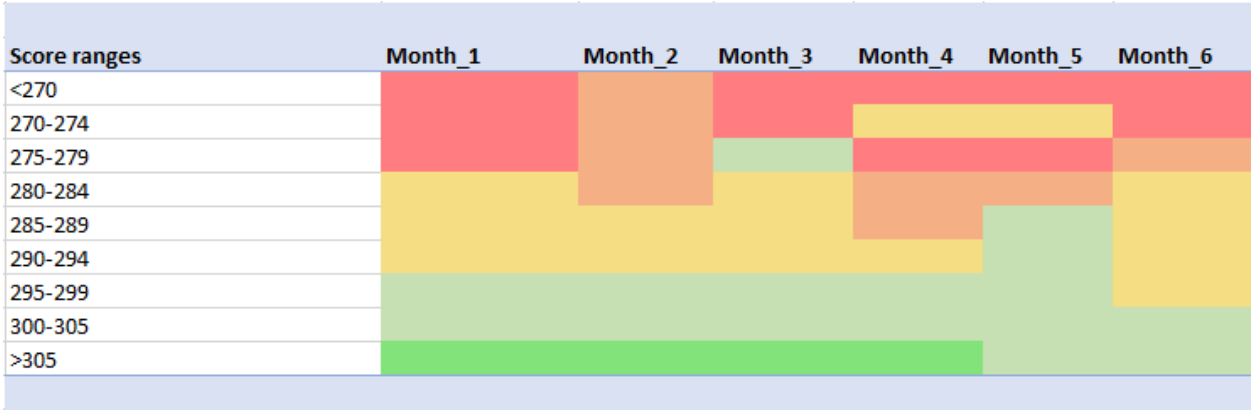


Figure 5. The stability of the final scoring card: the average probability of default.

Note: Colours are indicating the relative magnitude: the redder the cell the higher the number (meaning higher average probability of default), the greener the cell the smaller the number.

The ranking ability of my scoring card is relatively stable over time; colours are not changing randomly over several months. However, still, the scoring card needs constant monitoring after the implementation to ensure that the distribution of incoming population does not change significantly compared to the distribution of data the scorecard was constructed on (Siddiqi, 2005).

The next heat map (Figure 6) represents the ability of the developed scoring card to rank consistently surplus (monetary gain from the customer) magnitudes over time.

Score ranges	Month_1	Month_2	Month_3	Month_4	Month_5	Month_6
<270						
270-274						
275-279						
280-284						
285-289						
290-294						
295-299						
300-305						
>305						

Figure 6. The stability of the final scoring card – the magnitude of surplus.

Note: Colours are indicating the relative magnitude: the redder the cell the higher the number (higher surplus magnitudes), the darker blue the cell the smaller the number.

We can notice that the higher the score the darker red the cells are, meaning the more profitable are the loans that were issued. The dark blue cells are the groups of the loan applications that yield relatively small profit (or higher loss). Naturally, the organization or microfinance institution aims to minimize the share of those applications in the portfolio or to avoid having them in the portfolio at all.

6. Conclusions

In my master's thesis I used real data from a particular lending organization with 8268 observations, 13 independent variables and a dependent variable. I constructed my own dependent variable that is based on the surplus (in other words – monetary gain from the applicant) and defines whether the customer is "good" or "bad" in the following way:

- loan is "good" (the dependent variable takes the value of 0) if the customer paid the financial institution before the final deadline plus 30 days more than 100% of the loan principal amount;
- loan is "bad" (the dependent variable takes the value of 1) if the customer failed to pay the financial institution before the final deadline plus 30 days more than 100% of the loan principal amount.

Such dependent variable is focused on distinguishing profitable customers from the customers who are more likely to bring the loss for the financial organization. Therefore, the scoring card based on the new metric is expected to minimize the share of those customers in the loans' portfolio.

To build the predictive model, firstly, I used the machine learning algorithm called Random Forest to define which variables are the most useful among the other, this is important for giving higher weights for these variables in the final scoring card. The choice of the algorithm is based on earlier studies that showed that the Random Forest is able to build the most accurate classification model compared to other machine learning algorithms (Sharma, 2010). The paper also contributes to the emerging literature dedicated to the applications of machine learning algorithms to financial and credit scoring industry.

Next step I conducted the logistic regression to obtain the predictive probabilities of default for each characteristic. In my thesis I also described in detail the steps of the scoring card development and assessed the performance of my scoring card with Gini coefficient and the area under the ROC curve (AUROC). AUROC of my model is almost 0.7 (Lessmann et. al, 2015) - that is good for the model constructed on the real data. I also checked the stability of my model - it looks stable over time, that means it will not give unexpected results after the implementation unless the population changes.

In the master thesis, author showed that the scoring model, that is developed using a surplus-oriented dependent variable is able be a base for the stable scoring card that can rank the default quite clearly as well. The resulting model, that includes a mix of variables from different sources can be used in the microfinance institutions or banks to help the lenders make faster and more effective risk assessment of the incoming credit applications. In addition, it can be run in parallel as a second scorecard, to make more accurate decisions and enhance the power of existing scoring card.

As a further research would be interesting to see how the resulting model is able to increases the surplus for the financial organization after a couple of months after the implementation. As the model is based on the profit-oriented metric author would expect it to rank better than the model based on the "good"/"bad" dependent variable. Additionally, interesting to conduct the research about how other machine learning and artificial intelligence techniques can benefit the scoring card development process.

7. References

1. Galindo, Jorge & Tamayo, Pablo. (2000). Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications. Computational Economics. 15. 107-43. 10.1023/A:1008699112516.
2. Baker, M. J., & Churchill, G. A. (1977). The impact of physically attractive models on advertising evaluations. Journal of Marketing Research, 14(4), 538-555. http://dx.doi.org/10.2307/3151194
3. A. Steenackers, M.J. Goovaerts, A credit scoring model for personal loans, Insurance: Mathematics and Economics, Volume 8, Issue 1, 1989, Pages 31-34, ISSN 0167-6687, https://doi.org/10.1016/0167-6687(89)90044-9.
4. de Roure, Calebe and Pelizzon, Loriana and Tasca, Paolo, How Does P2P Lending Fit into the Consumer Credit Market? (April 20, 2016). Available at SSRN: https://ssrn.com/abstract=2756191or http://dx.doi.org/10.2139/ssrn.2756191
5. Desai VS, Conway DG, Crook JN, Overstreet GA. 1997. Credit scoring models in credit union environment using neural network and generic algorithms. IMA Journal of Mathematics Applied in Business & Industry 8: 323-346. https://doi.org/10.1093/imaman/8.4.323
6. S. Finlay, "Credit scoring for profitability objectives," European Journal of Operational Research, vol. 202, no. 2, pp. 528–537, 2010
7. Abdou, HAH and Pointon, J 2011, 'Credit scoring, statistical techniques and evaluation criteria: A review of the literature', Intelligent Systems in Accounting, Finance & Management, 18 (2-3), pp. 59-88.
8. Shweta Arya, Catherine Eckel, Colin Wichman, Anatomy of the credit score, Journal of Economic Behavior & Organization, Volume 95, 2013, Pages 175-185, ISSN 0167-2681.
9. Gang Dong, Kin Keung Lai, Jerome Yen, Credit scorecard based on logistic regression with random coefficients, Procedia Computer Science, Volume 1, Issue 1, 2010, Pages 2463-2468, ISSN 1877-0509.
10. Shu-Ting Luo, Bor-Wen Cheng, Chun-Hung Hsieh, Prediction model building with clustering-launched classification and support vector machines in credit scoring, Expert Systems with Applications, Volume 36, Issue 4, 2009, Pages 7562-7566, SSN 0957-4174.

11. Tian-Shyug Lee, Chih-Chou Chiu, Chi-Jie Lu, I-Fei Chen, Credit scoring using the hybrid neural discriminant technique, Expert Systems with Applications, Volume 23, Issue 3, 2002, Pages 245-254, ISSN 0957-4174, Hsieh, N.-C. Hybrid mining approach in the design of credit scoring models. Expert Syst. Appl. 2005, 28, 655–665.
12. Bates, J. M., and C. W. J. Granger. "The Combination of Forecasts." OR, vol. 20, no. 4, 1969, pp. 451–468. JSTOR, JSTOR.
13. Yao, P. Hybrid fuzzy SVM model using CART and MARS for credit scoring. In Proceedings of the International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC'09, Hangzhou, China, 26–27 August 2009; pp. 392–395.
14. Bijak, K.; Thomas, L.C. Does segmentation always improve model performance in credit scoring. Expert Syst. Appl. 2012, 39, 2433–2442.
15. Hand, D.J. and Henley, W.E. (1997) Statistical Classification Methods in Consumer Credit Scoring: A Review. Journal of Royal Statistical Society, 160, 523-541.
16. West, D. Neural network credit scoring models. Computers & Operations Research, 27 (11/12) pp. 1131-1152, 2000.
17. B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen. "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring." The Journal of the Operational Research Society 54, no. 6 (2003): 627-35. http://www.jstor.org/stable/4101754.Stavros A. Zenios. Practical financial optimization, Draft of July 22 2005.
18. Svetnik, Vladimir, Andy Liaw, Christopher Tong, J. Christopher Culberson, Robert P. Sheridan and Bradley P. Feuston. "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling." Journal of chemical information and computer sciences 43 6 (2003): 1947-58.
19. Breiman L (2001). "Random Forests". Machine Learning. 45 (1): 5-32. doi:10.1023/A:1010933404324
20. Liaw, Andy & Wiener, Matthew. (2001). Classification and Regression by RandomForest. Forest. 23.
21. Claude E. Shannon: A Mathematical Theory of Communication, Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, 1948.

22. Siddiqi, Naeem. 2006. Credit risk scorecards: developing and implementing intelligent credit scoring. Hoboken, N.J.: Wiley.
23. Park, Hyeoun-Ae. (2013). An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain. Journal of Korean Academy of Nursing. 43. 154-164. 10.4040/jkan.2013.43.2.154.
24. Allison, Paul David. 1999. Multiple regression: a primer. Thousand Oaks, Calif: Pine Forge Press.
25. Sheather, Simon J. 2009. A modern approach to regression with R. New York, NY: Springer.
26. Mays E., Lynas N. Credit Scoring for risk managers. Lexington, KY. 2017
27. Tabachnick, B. G., & Fidell, L. S. (1996). Using Multivariate Statistics (3rd ed.). New York: Harper Collins.
28. Chen, X. R. (2004). Gini coefficient and its estimation. Statistical Research, (8), 58-60
29. Apte C, Weiss S. 1997. Data Mining with Decision Trees and Decision Rules. Future Generation Computer Systems 13: 197-210.
30. Zhang J., Thomas C. Lyn. The effect of introducing economic variables into credit scorecards: an example from invoice discounting. volume 9, number 1 (March 2015) pp:57-78
31. Khashei, M.; Hamadani, A.Z.; Bijari, M. A novel hybrid classification model of artificial neural networks and multiple linear regression models. Expert Syst. Appl. 2012, 39, 2606–2620.
32. Andrew Foss, Osmar R. Zaiane, Estimating True And False Positive Rates In Higher Dimensional Problems and its Data Mining Applications. Data Mining Workshops, 2008
33. Bylander, T. Machine Learning (2002) 48: 287
34. Tarannum A. Bloch, Vaghela V.B., Wandra K.H., Applied Taxonomy Techniques Intended for Strenuous Random Forest Robustness, Tarannum A Bloch et al, Int. J. Comp. Tech. Appl., Vol 2 (6), 2061-2065
35. Sullivan, Hue & Christophe, Hurlin & Tokpavi, Sessi. (2017). Machine Learning for Credit Scoring: Improving Logistic Regression with Non Linear Decision Tree Effects.

36. Lessmann, Stefan & Baesens, Bart & Seow, Hsin-Vonn & Thomas, Lyn. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research.
37. N. Mandrekar J., Receiver Operating Characteristic Curve in Diagnostic Test Assessment Biostatistics for Clinicians Volume 5, Issue 9, September 2010, Pages 1315-1316
38. Allied Market Research. Global Peer to Peer Lending Market by End-user (Consumer Credit Loans, Small Business Loans, Student Loans, and Real Estate Loans) and Business Model Type (Alternate Marketplace Lending and Traditional Lending) - Global Opportunity Analysis and Industry Forecast, 2014-2022. Mar 2017
39. Michael A. Kuhn & Peter Kuhn & Marie Claire Villeval, 2014."Self-Control and Intertemporal Choice: Evidence from Glucose and Depletion Interventions," CESifo Working Paper Series 4609, CESifo Group Munich.
40. Daniel L. Chen Tobias J. Moskowitz Kelly Shue Decision Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires The Quarterly Journal of Economics, Volume 131, Issue 3, 1 August 2016, Pages 1181–1242.
41. John C. Hull (2009) The Credit Crunch of 2007: What Went Wrong? Why? What Lessons Can be Learned? The First Credit Market Turmoil of the 21st Century: pp. 161-174.
42. Hussein A. Abdou, John Pointon. Credit scoring, statistical techniques and evaluation criteria: a review of the literature, Volume18, Issue2-3 April-September 2011 Pages 59-88
43. Akbani R., Kwek S., Japkowicz N. (2004) Applying Support Vector Machines to Imbalanced Datasets. In: Boulicaut JF., Esposito F., Giannotti F., Pedreschi D. (eds) Machine Learning: ECML 2004. ECML 2004. Lecture Notes in Computer Science, vol 3201. Springer, Berlin, Heidelberg
44. Gastwirth, Joseph L. "The Estimation of the Lorenz Curve and Gini Index." The Review of Economics and Statistics 54, no. 3 (1972): 306-16. doi:10.2307/1937992.
45. Grömping, Ulrike. (2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest. The American Statistician. 63. 308-319. 10.1198/tast.2009.08199.
46. Caballero, Ricardo J., Emmanuel Farhi, and Pierre-Olivier Gourinchas. 2008. "An Equilibrium Model of "Global Imbalances" and Low Interest Rates." American Economic Review, 98 (1): 358-93.

47. Jorge Martínez Pagés & Antonio Millaruelo, 2016. "The recent application of negative policy interest rates in the euro area and in other economies: rationale and preliminary evidence on their effects," Economic Bulletin, Banco de España; Economic Bulletin Homepage, issue JUL, July.
48. Richard W. Cottle, Mukund N. Thapa Linear and Nonlinear Optimization Volume 253 of International Series in Operations Research & Management Science Springer, Jun 11, 2017

Non-exclusive licence to reproduce thesis and make thesis public

I, Kateryna Volkovska

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

Modelling the predictive performance of credit scoring by logistic regression and ensemble learning

supervised by Jaan Masso,

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, **24.05.2018**